STAT2

Module 1.1 Linear Regression

A **statistical** unit is one member of the set of entities being studied

A **population** is a collection of units about which research questions are asked

A **sample** is a subset of the population. Typically, samples should be representative

Inferential statistics and data science is the process of learning about relationships in a sample in a way that is reliable enough to generalize from the sample to a population of interest.

To **operationalize a concept** means to derive a set of steps to measure the concept

The **validity** of a dataset or measurement tool is the extent to which the dataset or measurement tool measures what it claims to measure

linear regression: Is used to explain or model the relationship between a single variable Y, and one or more variables $X_1, \ldots, X_p$

- Y is called the response, outcome, output, or dependent variable
- $X_1, \ldots, X_p$ are called predictors, inputs, independent variables, explanatory variables, or features. In some contexts, they are also called covariates.

Regression analysis has two main objectives:

- **Prediction**: predict an unmeasured/unseen Y using observed $X_1, ..., X_p$
- **Explanation**: To assess the effect of, or explain the relationship between, Y and $X_1, ..., X_p$.

$\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be the response variable and $\mathbf{x_1} = \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{pmatrix}, \mathbf{x_2} = \begin{pmatrix} x_{1,2} \\ x_{2,2} \\ \vdots \\ x_{n,2} \end{pmatrix}, \ldots, \mathbf{x_p} = \begin{pmatrix} x_{1,p} \\ x_{2,p} \\ \vdots \\ x_{n,p} \end{pmatrix}$ be predictors; we will collect the predictors in a matrix: $X = (1x_1x_2...x_p)$, where 1 = (1,1,. . . ,1)ˆT. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ be a vector of parameters. Finally, let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be a vector of error terms.

Definition/Assumptions of the linear regression model:
1. **Linearity** xian xing. X ~ Y scatter plot follows a Linear pattern.
2. **Independence** du li xing. Y is independent of errors/residuals.
3. **Homoskedasticity (constant variance)** fang cha qi xing. variance is the same for all X.
4. **Normality** zheng tai xing. residuals approximately normally distributed, with a mean of zero.

Interpreting simple linear regression parameters:
1. β_0: the intercept of the true regression line. β_0 is the average value of Y when x is zero. Usually this is called the "baseline average".
2. β_1: the slope of the true regression line. β_1: is the average change in Y associated with a 1-unit increase in the value of x.

Interpreting multiple linear regression parameters:
$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i(p-1)} + \varepsilon$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$F = kx$: F = Force; k = Spring Constant; x = Displacement

A **circular analysis or double dipping** is the process of exploring a dataset in an attempt to discover what relationships exist, and then test hypotheses related to that exploration on the same dataset.

Ways to avoid circular analyses:
1. Design the analysis and prespecify research hypotheses before observing the data.
2. Subset the data

Module 1.2 Least squares estimation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

: y is measured response; B_0 and B_1, E_i are unknown, to be estimated; X_i is measured predictors

The **line of best fit** to the data is the line that minimizes the sum of the squared vertical distances between the line y and the observed points

$y = X\beta + \varepsilon$: The problem is to find a B so that XB is as close as possible to y.

The **surface of best** fit to the data is the surface that minimizes the sum of the squared vertical distances between the surface and the observed points:

Let X be an $m \times n$ matrix, $\mathbf{v}$ be $n \times 1$, and $\mathbf{y}$ be $m \times 1$. Then:
1. Lemma 1: Then $X^T X$ is symmetric, i.e., $\left(X^T X\right)^T = X^T X$.
2. Lemma 2: Let $\mathbf{y} = X\mathbf{v}$. Then $\frac{\partial y}{\partial \mathbf{v}} = X$ and $\frac{\partial y^T}{\partial \mathbf{v}} = X^T$
3. Lemma 3: Let $c = \mathbf{v}^T \left(X^T X\right) \mathbf{v}$. Then $\frac{\partial c}{\partial \mathbf{v}} = 2X^T X\mathbf{v}$

The **residuals** are defined as:

The **fitted values** are defined as:

The **hat matrix**, H, is defined as:

Least Squares Estimation: We define the best estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ as the one that minimized the sum of the squared residuals:

In order to use least squares, we assume that: 1. $E(\varepsilon_i) = 0$ for all $i = 1, \ldots, n$.
2. $E(Y_i) = \mathbf{x_i}^T \boldsymbol{\beta}$ for all $i = 1, \ldots, n$.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$
4. $\left(X^T X\right)^{-1}$ exists.

The **Gauss-Markov Theorem**: Suppose that:
1.E $(\varepsilon_i) = 0$ for all $i = 1, \ldots, n$.
2. $E(Y_i) = \mathbf{x_i}^T \boldsymbol{\beta}$ for all $i = 1, \ldots, n$.
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$
4. $\left(X^T X\right)^{-1}$ exists.
Then $\widehat{\boldsymbol{\beta}}$ is the "best" unbiased estimator of $\boldsymbol{\beta}$.

The **maximum likelihood estimator**.
Suppose that $\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$. Then:
1. marginal pdf:
2. joint pdf:
3. log-likelihood:

Sums of squares:
1. RSS: Residual sum of squares:
2. ESS: Explained (or regression) sum of squares:
3. TSS: Total sum of squares:

The residual sum of squares **RSS** can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.

The parameter σ^2 determines the amount of spread about the true regression line.

An estimate of σ^2 will be used in statistical inference (e.g., confidence interval formulas and hypothesis testing), presented in the next two sections.

Note:
1. The divisor $n-(p+1)$ in is the number of degrees of freedom (df) associated with RSS and $\hat{\sigma}^2$.
2. The RSS has $n-(p+1)$ df because $p+1$ parameters must first be estimated to compute it, which results in a loss of $p+1$df.
3. Replacing each y_i in the formula for $\widehat{\sigma}^2$ by the r.v. Y_i gives a random variable.
4. It can be shown that the r.v. $\hat{\sigma}^2$ is an unbiased estimator for $\widehat{\sigma}^2$.

The coefficient of determination, R^2, is defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Note: - $0 \leq R^2 \leq 1$
- Assuming that the model is correct, R^2 is interpreted as the proportion of observed variation in y explained by the model.

Warnings about R^2
1. R^2 can be close to 1 but the model is the wrong fit for the data.
2. R^2 can be close to 0 even when the model is the correct fit for the data.
3. R^2 should not be used to compare models with a different number of predictors.
4. R^2 says nothing about the causal relationship between the predictors and the response.

The least squares estimate is the solution to the normal equations:

$$X^T X\boldsymbol{\beta} = X^T\mathbf{Y}.$$

1. When $\left(X^T X\right)^{-1}$ exists, there is a unique solution, $\widehat{\boldsymbol{\beta}}$.
2. When $\left(X^T X\right)^{-1}$ does not exist, there will be infinitely many solutions.

Definition: When $\left(X^T X\right)^{-1}$ does not exist, the regression model is said to be non-identifiable (or, unidentifiable).

Why might we have non-identifiability?
1. One variable is just a multiple of another.
2. One variable is a linear combination of several others.
3. There are more variables than members in the sample.
Note: Near non-identifiability is trickier than exact non-identifiability.

OLS (Ord. least Squares):
For simple, $\varepsilon_i = y_i - E(y_i) = y_i - (\beta_0 + \beta_1 x_i)$
$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$
$\frac{\partial Q}{\partial \beta_0} = 2\sum(y_i - \beta_0 - \beta_1 x_i)(-1) \equiv 0$